Encoding Race and Structural Racism: Philosophical Perspectives

Arden Ali, arali@clarku.edu

Sally Haslanger, shaslang@mit.edu

Jerome Hodges, jh4@jainfamilyinstitute.org

Lily Hu, lilyhu@g.harvard.edu

Abstract

What is race? In what sense is race "real"; in what sense is it an illusion? How should scientific inquiry take race into account (or not)? How should systems that distribute key social benefits and burdens take race into account (or not)? Does fairness require colorblindness? If not, how can we incorporate considerations of race that do not lead to other kinds of harms? These questions arise in a broad range of disciplines, but have now taken on another dimension of significance as automated computational and machine learning tools are increasingly folded into systems that address complex social coordination problems. After reviewing the philosophical literature, we argue that the main issue is not how or whether we adequately *represent* racial phenomena, or how we can trade-off legitimate "non-race-related" interests with an aim of racial "fairness," but how such computational systems contribute to the *production* of race; the normative issue, then, is not how to be fair, but how to dismantle racism.

1. Introduction

Over the past decade, the burgeoning field of "algorithmic fairness," also commonly known as "fair machine learning," has been concerned with the problem of *bias* in data-based predictive systems. Computation and machine learning based tools have increasingly been incorporated into decision-making processes that distribute key social benefits and burdens ranging from employment opportunities to second chances at public life. There is evidence, however, that algorithms have introduced, reproduced, and exacerbated troubling social biases in high-stakes social domains. Most prominent are concerns regarding the potential harms that data-based tools perpetrate along axes of salient social categories such as race and gender. Naturally then, one particularly active line of scholarly inquiry focuses on identifying and proposing approaches to resolve cases of "algorithmic discrimination."

One primary task of this cross-disciplinary effort has been to fill in the content of the core ethical notions of "discrimination" and "fairness" in a way that illuminates the socially and normatively salient features of the problems that animate concerns about algorithms and bias. Though a substantial part of the scholarship on algorithmic fairness has focused on situating algorithms and their effects within various legal regimes, the law does not offer clear guidance on what constitutes "algorithmic discrimination"

interpreted even within the narrow frames of anti-discrimination doctrine.[1] Neither do proposals for achieving algorithmic racial "fairness" make explicit connection to philosophical theories of race nor endorse particular substantive theories of what constitutes racial justice, making it often unclear what the normative justifications for the various classes of approaches exactly are. This presents a basic hurdle to progress, as it leaves opaque what grounds there are for preferring one class of approaches to achieving algorithmic fairness to others.

The primary aim of this paper is to bridge the gap between proposed technical definitions of and solutions to the problem of algorithmic racial discrimination and the normative assumptions guiding efforts to promote racial justice in complex sociotechnical systems. To that end, we will survey the dominant schools of thought in the literature on what constitutes algorithmic "fairness" or "non-discrimination" and probe their underlying theories of race, racism, and fairness.

In Section 2, we will introduce the main philosophical approaches to race and racism. Then, in Section 3, we provide an overview of prevailing technical approaches towards fair computational systems and draw out their connections to the previously discussed theories of race and racism, with the aim of clarifying the strengths and weaknesses of recent anti-discrimination strategies. Our analysis reveals a set of background individualistic assumptions undergirding these frameworks that are inadequate to address the problems of systemic injustice. In Section 4, we consider in more detail why these individualist strategies are inadequate. We argue that although predominant approaches to fairness in computational systems rightly assumes that justice requires "treating likes alike, and unlikes unlike," a crucial challenge for any use of this principle lies in determining what are the morally relevant similarities and differences that legitimately count as a basis for treatment. In considering race and racism, the operative question becomes: When is race morally relevant and when not? What is at issue in considering the moral relevance of race? (And in the background: What is race, anyway?) We argue further that a normative focus on "fair treatment" tends to obscure how all "treatment" is embedded in structures that distribute power and vulnerability, and concern about whether we should treat individuals "the same" or "differently" *within* a structure deflects attention from the injustice of the structure *itself*. Unjust social structures, not only discrimination in narrowly defined decision-making processes, should be among the targets of antiracist efforts. We maintain that those working in machine learning would benefit considerably in their antiracist efforts if they worked collaboratively with social and political theorists who are in a position to identify the social processes by which race and racism have been and continue to be produced, and with moral philosophers who can help with the work of articulating normative principles—beyond fair treatment—to identify and intervene in racist structures.

[1] Many scholars have written on the cluster of thorny issues concerning how racial impacts of algorithms fit within legal doctrine. By no means an exhaustive list, see, e.g., Barocas & Selbst 2016, Kroll et al. 2016, Kim 2017, Huq 2019, Mayson 2019, Hellman 2020.

2. Is Race "Real" or Not?

In the 20th century, the discussion of race in philosophy largely focused on whether or not race is a natural kind, or a "real" division among humans.[2] This was taken to be relevant to normative questions about racism, for historically, beliefs about racial essences have undergirded racial hierarchy. If race is not "real," then treating members of different purported "races" differently would be based on a false presupposition. Most philosophers concluded that because races are not genuine natural kinds - because race is an illusion - we should pursue colorblindness as a normative ideal.

The strategy of claiming that race is an illusion depends on the idea that the English term 'race' includes as part of its meaning that race is a biological category. Evidence for this are the widespread beliefs that race is inherited and racial "passing" is possible; regardless of one's social identity, one's race is fixed by blood relations. Very generally, those who maintain that 'race' is applied correctly only if physical—or more specifically biological—conditions are met fall into four camps. Let's start with the now discredited view as the contrast to which other views are positioned:

> ***Racialism:*** Races exist and are natural kinds. One is born into one (and only one) of these kinds and membership in the kind determines, due to the kind's essence, a wide range of one's physical, psychological, and moral characteristics. Racial segregation, stratification, etc. is warranted, due to the natures of the different races.

The flat-out rejection of this view holds that because there is no such natural basis for our racial categories, race is an illusion:

> ***Anti-Realism:*** (Appiah 1996; Blum 2002; Glasgow 2009) There are no biologically-based differences between what we ordinarily consider to be racial groups. Because the concept of race presumes a biological basis, races do not exist. Treating people differently *based on race*, is misguided. Most anti-realists in philosophy argue, however, that racialized groups—groups *presumed* to be distinguishable as biologically distinct races and treated as such—do exist. It is an open question how we should treat racialized groups.

Not all philosophers are willing to accept the claim that there are *no* biologically-based differences between races. So another approach maintains that there are ways of accommodating the idea that races exhibit biological differences, and so might count as biologically "real" human groups, without accepting the racialist's problematic moral conclusions. Call these views *minimalist accounts* of race:

> ***Populationist Racial Realism:*** (Spencer 2014, 2015) Races exist and can be characterized genetically. Population genetics has discovered that humans can be divided into five groups based on genetic traits, roughly corresponding to continental geography (Risch et al 2002; Rosenberg et al 2002; cf. Bolnick 2008). These five groups also map onto the ordinary divisions

[2] Historically this claim has been closely associated with a biologistic conception of race; but even before the study of human genetics, racial kinds were taken to be grounded in metaphysics or theology (Stocking 1994).

we make between White, Black, Asian, Indigenous North/South American, Pacific Islander. According to the populationist race account, races are "groups of human beings whose visible physical differences are *genetically transmitted* and indicative of distinct *founding populations* that were once *reproductively isolated* from one another due to geographical separation." (Zheng 2018; Hardimon 2017, 3) On this view, the genetic differences do not (or do not currently seem to) make a substantial difference to the dispositions, capacities, or behavior of those in the groups. As a result, this view retains a biological realism about race but avoids many of the moral and metaphysical problems with racialism.

Minimal Racial Realism: (Hardimon 2017, 2019) Races are groups of human beings distinguishable by visible physical traits that signal differences in geographical ancestry. The visible physical traits that characterize a group may or may not be based in a particular genomic signature. But, it is claimed, differences in physical traits (skin color, eye shape, hair texture) among racial groups are biologically caused.

These minimalist accounts of race retain a biological realism about race but avoid many of the moral and metaphysical problems with racialism, e.g., they reject a commitment to racial essences that determine morally relevant features of different races. The supposed metaphysical or biological grounding of differential treatment disappears. As a result, most minimalists argue that race, for the most part, should not be taken into account in social practices and policies.[3]

The approaches described so far tend to share three background ideas.

1. **Individualist Race:** Race, if it exists, is an inherited, intrinsic[4], and ineradicable feature of individuals. One's membership in a race does not vary over time or context; it is not the result of choice, culture, or identity.

2. **Aristotelian Principle:** The underlying moral principle used to connect the metaphysics of race with practice is that we should treat likes alike, and unlikes unlike.[5] More specifically, morally irrelevant differences should not be a basis for differential treatment.

3. **Individualist Racism:** Racism is a harm or wrong done *to* individuals *by* individuals or organizations. The mechanism of racism is interactional; *how agents treat other agents* . One example is an **attitudinal** view of racism, which takes racism to occur when individuals or organizations

[3] Spencer argues that populationist race could be medically relevant and should be considered in the context of medical research.

[4] Within the philosophical literature, to say that race is intrinsic is to say that one has a race simply by virtue of one's non-relational features; one would have that race regardless of social context just by virtue of one's individual properties. (Compare, x is longer than y (relational), and x is spherical (non-relational).) According to this usage, intrinsicness does not entail anything about essence or necessity.

[5] In the "algorithmic fairness" literature that we will discuss in greater detail in the following section, this approach reflects what Friedler, Scheidegger, and Venkatasubramanian (2021), following Dwork et al. (2012), call an "individual fairness" approach to race in computational systems.

> treat irrelevant or merely imagined "racial" differences between people as a basis for different treatment. Individual racism an example of a violation of the Aristotelian Principle: people of different races are treated differently on the basis of morally irrelevant differences.

In what follows, we will first consider approaches to algorithmic bias that seem to accept these assumptions. We will argue that although there are circumstances when the proposed remedies are apt, such approaches are too limited because these aforementioned accounts present inadequate analyses of race and racism. In some cases, we should attend to and disrupt individual racism, but this is not enough. Our own proposals are motivated by a different understanding of race and racism based on the following concerns:

Re 1 **(Race as Relational)**: Given the overwhelming evidence against racialism (the view that race is a meaningful biological division between humans that warrants differential treatment), scholars have developed a new strategy for thinking about race. Race may not be *biologically* meaningful, but it is certainly *socially* meaningful. In other words, the groups we count as races have been, and continue to be, viewed and treated in ways that situate them in a social hierarchy. Moreover, racial categorization differs in different cultural and historical contexts.

This suggests that races are social kinds rather than biological kinds. One has a race by virtue of the social relations and social structures one occupies: like being a teacher or a student, a landlord or a tenant, race is a relational rather than intrinsic property of individuals. (See, e.g., Haslanger 2000). For example, on a view of this sort, for an individual to be White is for them to be privileged in a system of racial hierarchy based on physical markings - such as fair skin, eyes without epicanthal folds, hair that is not tightly curled - that are assumed to link one to a time and place of ancestral origin. The focus of inquiry, then is *not* whether there are underlying biological differences between the races. The question becomes: What are the social processes that create racial categorization and stratification?

Re 2 **(Moral Relevance)**: Although morally irrelevant differences should not be a basis for differential treatment, the Aristotelian principle, as stated, does not elaborate what differences are morally relevant and in what contexts. For example, in the case of gender, there are contexts in which women and men should be treated the same because there are no relevant moral differences, e.g., women and men should earn equal pay for equal work. However, there are other contexts in which the differences between women and men are relevant, e.g., in certain medical contexts. Moreover, the principle seems to assume that what is relevantly "like" and "unlike" is a fact independent of the treatment in question. But how people are treated can *cause* morally relevant differences.

For example, standardized tests such as the SAT treat all students the same in the sense that all students answer the same questions when sitting for the exam. However, individuals taking the test come from different backgrounds and have different vocabularies. On the face of it, differences in vocabulary are not relevant in determining intelligence or capacity to excel in college; whether one knows the parts of a yacht or can name the different agencies falling within the social services system in one's community shouldn't matter. But the SAT test may result in different scores that track knowledge of vocabulary (such as the parts of a yacht) that is irrelevant to success in college. That two individuals have different vocabularies is

not morally relevant, but it has a moral impact because the test uses it to create morally relevant access to goods.

Re 3 (**Structural Racism**): If we allow that race is an imposed social status rather than a morally irrelevant feature of individuals, then antiracist efforts should not center on aspirations toward colorblindness. The goal should instead be to *undo race*, i.e., to disrupt the processes that create and sustain race as we know it. Of course some of this will involve changing the minds and actions of racist agents—individuals, organizations, and such—but we are all implicated in the broader social practices and structures that produce and maintain race and racial injustice. In treating "likes alike" and "unlikes unlike", antiracism must be nimble and alert: what are the morally relevant similarities and differences *in this context*? And are we participating in the production of race or *can we actively counter it*?

There is a crucial contrast here: traditional conceptions of race and racism focus on individuals, e.g., whether an individual is a member of a race by virtue of their biology, and whether an action is racist because it is caused by bad attitudes. The alternative structural conceptions of race and racism we have started to sketch instead concentrate on social relations and networks of relations. Structural accounts do not deny that there may be underlying biological differences and bad attitudes; but the issues of racial justice are broader. We should be asking: how do observed or imagined "color" differences[6] result in social stratification? What are the practices, norms, laws, and material conditions that produce and reproduce racial hierarchy, even by those with good intentions?

In the following section, we show that up to now, dominant frameworks toward fairness in computational systems have not been primarily concerned with these social-structural accounts but are instead undergirded by a more individualistic orientation towards race and racism. In presenting overviews of the range of strategies that fall under this umbrella of *race neutrality-based approaches*, we will show that each class of these approaches is rationalized by the thought that what makes a system liable to being racially unfair is its tendency to "act on" racial information or allow race to influence its operations in an illicit way.

3. Race-Neutrality Approaches

The class of race-neutrality approaches toward fairness are motivated by a concern that overt, proxy-based, or indirect discrimination are the primary avenues by which racial groups can be disadvantaged by the use of data-based computational systems, and that efforts at addressing unfairness in these systems should be focused on eliminating these kinds of harms. We survey in this section these dominant race-neutrality strategies toward algorithmic fairness, which we group into three classes: (1) data-driven, (2) statistical parity-based, and (3) causal and counterfactual. We hope to show that themes of individualism about race and racism, as well as Aristotelian conceptions of fairness, are embedded deeply in each of these broad approaches to the issue at hand.

[6] We use the term "color" (in scare quotes) to refer to the locally salient set of racial markings, which may or may not prioritize skin color. Racialization works differently in different contexts (cf. the United States, South Africa, Brazil), and correspondingly, the markers of race differ as well.

3A. *Race-Neutrality Approach I: Data-Driven*

Data-driven approaches to algorithmic fairness, while employing a variety of technical methods, share an inclination to address fairness concerns through the selection or manipulation of the data that are used in training and/or executing the machine learning system. The goal of this section is to show that individualism about racism and the Aristotelian principle undergirds these methods.

The most flat-footed data-driven approach to racial fairness looks to exclude the "race variable" from the system entirely. This tactic follows rather straightforwardly from the worry that algorithms that incorporate the variable in its computations engage in explicit racial categorization. Inclusion in the system's model, the worry goes, amounts to directly penalizing or advantaging individuals based on their race. This approach is typified by the "Fairness Through Unawareness" criterion discussed in the computer science literature on fair machine learning (e.g. Gajane & Pechenizkiy 2017; Kusner et al. 2017; Mitchell et al. 2018), and is also supported by many legal scholars who worry that a governmental predictive system's use of racial information "directly" would unconstitutionally violate the 14th amendment's Equal Protection Clause,[7] or that employer use of such systems might run afoul of the disparate treatment prohibitions of Title VII.[8]

The reasoning is simple enough: if we can guarantee that a machine learning system doesn't have access to the race of its decision subjects, i.e., if we can ensure that the system is rid of racial information, we can guarantee that it isn't using race in a suspect way. Accordingly, ensuring that the system is thoroughly race-neutral—in its inputs, learning process, and as a result, its outputs—will in turn ensure its racial fairness. As Chief Justice Roberts pithily put it, "the way to stop discrimination on the basis of race is to stop discriminating on the basis of race."[9]

The unawareness approach embodies both the Aristotelian conception of fairness and the individualistic conception of racism that underpins it. The approach is Aristotelian because it tries to achieve fairness by removing the effect of a morally-irrelevant factor, namely race, on the system's decisions. It is individualistic in how it looks to achieve the Aristotelian principle: it seeks to excise illicit racial effects by manipulating the racial information available to the system so that it cannot treat individuals differentially on the basis of race. If the machine learning system is not trained on data including the race variable, and if it remains ignorant about the racial identity of its decision subjects, there is no obvious way that any decision subject could be targeted for harmful treatment. And if racism requires the targeting of individuals for differential treatment in virtue of their race, as the individualist has it, there is no way for an unaware algorithm to be racially unfair.

---

[7] Though many legal scholars, policymakers, and racial justice advocates *do* call for the elimination of the use of these variables entirely from a system's functioning and cite anti-discrimination doctrine as requiring this sort of exclusion, it is a matter of ongoing legal debate whether the law does require this sort of formalistic "colorblinding" approach. For differing legal interpretations on the issue—whether and to what extent anti-discrimination law requires "race-neutrality" in the data that predictive algorithms draw on—see e.g., Starr (2014), Mayson (2019), Yang & Dobbie (2020), Hellman (2020).

[8] See, e.g., Barocas and Selbst (2016).

[9] *Parents Involved in Community Schools v. Seattle* (551 U.S. 701 (2007))

One problem with "Fairness Through Unawareness" is that race has already shaped our world in material ways. This means that race is embedded not only in the race column of a given dataset, but in many other features of the social world. If this is true, fairness cannot be achieved merely by blocking observation of racial categories. In a city which is *in fact* perfectly racially segregated — with all White people living on one side of the tracks, and all people of color living on the other — an agent or system can be certain about the race of a subject by observing their "non-race" attribute (in this case, neighborhood of residence).

Once one appreciates this fact, it seems one must not only block the use of racial data, but also take care to address the ways that race can be inferred from *proxies* for race. A classic example of proxy discrimination is given in the long history of redlining (which continues to this day): the use of zip code as a way to target the exclusion of Black communities from receiving federal mortgage support. The correlations between zip code and race make it possible for such policies to appear "race-neutral"—i.e., not based on express usage of racial category membership—while still achieving agencies' desired effect of racial exclusion. Thus, algorithms that draw on zip code information might similarly engage in spatial discrimination generating outcomes that systematically track the outcomes that would be produced as a result of overt racial discrimination.

Those who acknowledge that proxies present a hurdle to ensuring that computational systems do not draw on racial information often instead seek to achieve 'colorblindness' instead of mere unawareness. To truly achieve a colorblind dataset, and thus a racially fair machine learning system, one must not only remove the race variable from a system's operations, but also proxies for race, which may include features that track social factors ranging from neighborhood information to past engagements with police.

Notice, however, that the motivation behind this colorblind strategy remains basically the same as the unawareness strategy. In trying to achieve a colorblind dataset, one is trying to make it impossible for a system to render a verdict on the basis of the decision subject's race. After all, if the system has been freed of racial information, including proxies for race, then it cannot target individuals in virtue of their race, and thus, there is no way for the system to be racially unfair. Thus, even while a broader range of features is scrutinized in an approach seeking colorblindness through eliminating proxies, the underlying reasons for pursuing the method are the same: to be racially fair, a system must not be able to glean any racial information, lest it act on race in its treatment of individuals.

Even more sophisticated efforts, which go beyond unawareness or colorblindness, toward achieving racially-neutral datasets show some commitment to individualism. One such data-driven route to racial fairness is the exclusion of 'dirty' datasets from machine learning systems entirely. The term 'dirty data' was traditionally used to refer to a multitude of ways that a dataset may be inaccurate, incomplete, or otherwise imperfect in a technical sense (Kim et. al. 2003). A dirty dataset is one that features missing data, unusable data, or errors in data entry. In some scholars' analyses, the category has been expanded to include datasets that are "derived from or influenced by corrupt, biased, and unlawful practices" (Richardson et al. 2019). This extended definition now means to include the kind of biased dataset that results from years of manipulated or 'juked' crime statistics and corrupt police practices, and which might be used to train a predictive policing algorithm (Richardson et al. 2019).

Undoubtedly, the historical practices that have produced data should be subject to special scrutiny. Even so, there is a commitment to individualism underpinning the idea that racial fairness can be achieved by (merely) paying attention to the role of dirty data in machine learning. On this line of reasoning, the kinds of datasets that are excluded are those that result from practices that trace to racial antipathy, ill will, or failures of respect. It is a clear demonstration of racial antipathy on the part of the police to plant drugs on innocent people, use unnecessary force in communities of color, and engage in racial profiling. The dataset is rejected because it traces to such practices. This means that the propensity to focus on dirty data as a means of achieving racial fairness may also encode a commitment to the idea that the harms of racism must derive in some way from the attitudes of wrongdoers. This is one manifestation of individualistic modes of thinking about racism.

If one is not interested in eliminating data or datasets altogether, a last data-driven strategy is to manipulate the data in order to debias it. Consider one simple method to 'clean up' data in this way: augmentation. Suppose we are looking at a dataset that includes information about individuals employed by a hospital, and we are trying to design a system for human resources to autofill forms based on limited demographic information. The data includes information about the hospital workers' race and occupation at the institution. You notice that the data is racially biased. Individuals whose racial identity is listed as Black are disproportionately assigned to service and support roles, while those whose identity is listed as White tend to be assigned to clinical and administrative roles. To 'debias' the dataset, one could augment the data by adding synthetic data points that counteract the skew. If we add a synthetic case of a Black individual in a clinical role, or a White individual in a service role, one can imagine slowly 'correcting' the bias in the dataset. When the data reaches the ideal non-biased state, there is no further need to add synthetic data points. One can then train an algorithm on this new dataset, which includes both the original data and the synthetic data.

The augmentation technique just described is rudimentary. But even sophisticated pre-processing techniques share the same theoretical orientation. Take, for example, the use of generative adversarial networks (GANs) to debias datasets. In these systems, you train a generative subsystem to create synthetic data points that a discriminator subsystem is unable to distinguish from the real data points. Specific GANs have been shown to generate indistinguishable synthetic data that achieves fairness goals (Sattigeri et. al, 2019).

The augmentation technique, and related data manipulation strategies, all incorporate familiar themes of Aristotelian fairness and individualism. The overall approach to fairness is to ensure that a morally-irrelevant difference (i.e. race) is exerting no influence on the decisions of the system. By manipulating the data until the racial bias dissipates, one seeks to render the system effectively ignorant about race. While the unawareness and colorblindness strategies reached this result by removing information, augmentation strategies get there by diluting the existing data until race has no predictive value. In the end, the conceptual underpinning is the same. If augmentation renders race predictively useless, and the harms of racism must be targeted at an individual in virtue of their race, then the system cannot be racially unfair in its predictions. After all, its predictions cannot be influenced by a factor with no predictive value.

3B. *Race-Neutrality Approach II: Statistical Parity-Constraint-based Approaches*

A second class of approaches to ensuring fairness looks to the statistical properties of a computational system's model and/or decisions. These "statistical criteria of fairness"[10] tie racial fairness to racial *parity*[11] with respect to some specified statistical measure. They differ among one another with respect to *which* statistical measure they take to be constitutive of fairness.

Parity-based conceptualizations of fairness must be tailored to the technical particularities of the computational system in question, and a long menu of metrics have been proposed in the algorithmic fairness literature. For illustrative purposes, we run through a sampling of these approaches with the aim of highlighting the principles of fairness that undergird them.

A simple system might be meant to predict some binary attribute Y which can be 0 or 1. Given a subject's features, it outputs its guess D: 1 if it predicts the subject has $Y = 1$, 0 if it predicts the subject has $Y = 0$. For example, a computational system might be tasked with predicting whether an individual will successfully repay a loan if offered. It will base its model and output on historical data of individuals and their eventual loan outcomes (whether they repaid successfully or defaulted). An individual who *in fact* repaid successfully has Y = 1; one who *in fact* defaulted has Y= 0. An individual *predicted* to successfully repay receives an assignment D = 1; one who is *predicted* to default receives an assignment D = 0.

For such a predictor we can ask such questions as:[12]

1. How many subjects are assigned $D = 1$?
2. How likely is it that the predictor will guess correctly?
3. How likely is it that the predictor will (incorrectly) assign $D = 0$ for a subject with $Y = 1$?
4. How likely is it that the predictor will (incorrectly) assign $D = 1$ for a subject with $Y = 0$?
5. How likely is it that a subject assigned $D = 0$ in fact has $Y = 0$?
6. What is the ratio of (3) to (4)?
7. etc.

For each such measure, we can ask: is this metric balanced, or close to being balanced, across racial[13] groups? Demanding that it be so amounts to imposing a statistical fairness constraint on the system. Different kinds of these have been given different names, including *Demographic Parity*, *Equal Accuracy*, *Equality of False Positive (or Negative) Rates*, *Equal Treatment Ratios,* and so on.

More complex decision systems offer more fine-grained output than the simple D = 1 or 0 of a binary classifier; they output instead a "score" R, typically a number between 0 and 1, where a subject's score is

[10] Hedden's (2021) p. 214.

[11] "Parity" has sometimes been used as a name for statistical balance with respect to a particular property, discussed below; here we use it in a more general sense, to refer to statistical balance with respect to any specified property.

[12] See Mitchell et al. for a much more exhaustive list.

[13] We focus on race, but of course the constraints are at least coherent (whether or not they have any relationship to pretheoretic conceptions of fairness) for any way of grouping decision subjects. One can enforce parity constraints across groups with individuals whose favorite color is blue vs. red.

usually understood as an estimate of the probability that the subject has $Y = 1$. To continue with the example of loan repayment prediction, the system could output a score corresponding to the probability that the individual will repay their loan successfully. These systems admit of more fine-grained questions, such as:

8. What is the average risk score R?
9. What is the average risk score R for subjects with $Y = 1$?
10. For a given risk score R, how many people are assigned R?
11. For a given risk score R, what fraction of people with $Y = 1$ are assigned R?
12. etc.

As before, for each such measure we can ask: is this the same across racial groups? And again, the corresponding constraints requiring equality of the measure across groups have been given different names, such as *Calibration*, *Equalized Odds*, and *Balance for False Positives*.[14]

Constraints of these and similar forms are all parity constraints: they require that some summary statistic of the decision process be the same across groups. These constraints are operationalizations of the thought that fairness or nondiscrimination requires some form of *balance* in how the system treats members of different racial groups.

Let's now consider what applying parity constraints to our computational system that looks to distribute loans according to data-based algorithmic predictions of successful repayment comes to. A fairness metric requiring balance with respect to criterion (1) suggests that fairness requires that the system judge that White and Black borrowers are equally likely to repay. A metric based on (2) suggests that fairness requires that the system not be less accurate in its predictions for Black borrowers than White borrowers, or vice versa. One based on (3) suggests that fairness requires that among individuals who will "in fact" repay their loan, the probability of being judged creditworthy by the system is the same whether one is White or Black.

At first blush, these criteria might *all* seem plausible as interpretations of what fairness requires in a loan decision system. So why not simply build an algorithmic system that satisfies them all? Unfortunately, we can't: it's a mathematical fact that so long as the distribution of the predicted variable is unequal across the relevant groups, no predictor, that is not trivial and not perfect, simultaneously satisfies all parity

[14] For a somewhat tongue-in-cheek overview and critique of the explosion of "fairness definitions" that have so far been introduced into the technical algorithmic fairness literature, see Arvind Narayanan, "Translation Tutorial: 21 Fairness Definitions and Their Politics," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2018. Carmelia Simoiu, Sam Corbett-Davies, and Sharad Goel argue that many of the parity-based definitions provide misleading representations of a system's fairness because of the problem of inframarginality in "The Problem of Infra-marginality in Outcome Tests for Discrimination," https://arxiv.org/pdf/1607.05376.pdf. Corbett-Davis and Goel levy an even broader set of critiques in "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," https://arxiv.org/pdf/1808.00023.pdf.

metrics that intuitively seem desirable in a system's functioning.[15] The upshot then is that the various statistical parity "definitions" of fairness that have been proposed encode *competing* conceptions of fairness.

Since they can't all be right, how do we decide among them? What underlying judgments might drive commitment to one or another parity-based approach? The proposals tend to doggedly avoid such questions.

To get clearer about the proposals, two questions need answering. First: what, exactly, is the proposed relationship between parity and fairness? There are at least three options:

> Parity is *constitutive* of fairness: violations of parity are *per se* unfair: for a system to be unfair *just is* for it to violate that notion of parity.

(The mathematical impossibility results have often been taken to suggest that the fairness in algorithms debate now bottoms out at a choice about which constraint expresses what it is for a system to be fair or nondiscriminatory.)

> Parity is *necessary and sufficient* for fairness: for a system to be unfair, it is necessary and sufficient that it violate the parity requirement, whether or not this requirement is definitional of fairness.[16]

> Parity is *evidence* of fairness: if a system satisfies the parity requirement, there is good reason to believe that the system is fair; if not, there is good reason to believe it is unfair.

Different answers to this first question lead to very different proposals. A statistical metric that successfully *defined* fairness would settle most algorithmic bias questions for good. A statistical metric that merely claims an evidential link between a statistical property and fairness, on the other hand, raises further questions about when the evidence might mislead.

The second question needing an answer is similar: to what, exactly, is the proposed constraint meant to apply? When asking whether a constraint holds, we might be asking one of two different questions:

---

[15] Different versions of this result apply to different problem setups. See Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. 2018. http://www.fairmlbook.org; Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2):153–163, 2017; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. 2016. https://arxiv.org/abs/1609.05807; Thomas Miconi, The impossibility of "fairness": a generalized impossibility result for decisions. 2017. https://arxiv.org/abs/1707.01195

[16] For a condition to be necessary and sufficient for fairness means: whenever the condition holds, fairness holds, and *vice versa*. Or sometimes it's used to mean, more strongly: it's impossible for an unfair system to satisfy the condition, and impossible for a fair system to fail to satisfy it. The argument below applies to either notion. These are strictly weaker than the condition *defining*, or *constituting*, or *being conceptually the same as* fairness. Triangles are polygons with three angles; call a polygon whose interior angles sum to π a "π-angle". There is a clear difference between the two concepts, even though they're (mathematically) necessary and sufficient for one another. Importantly, "necessary and sufficient" is often used relative to a tacit or background set of restrictions on the space of possibilities: being a triangle is necessary and sufficient for being a π-angle — but only in a flat space (ignoring spherical triangles).

> *The observational* question: is the relevant statistical property[17] *in fact* equal (or reasonably close to equal) across racial groups? Answering this question generally does not require knowing about the workings of an algorithmic predictor. So long as we know the distribution of predictions across groups, and (for constraints that require it) the distribution of the true values of the property being predicted across groups, we are in a position to evaluate whether the system satisfies the constraint.
>
> *The system question*: probabilistically, does such a system *tend to* balance the relevant statistical property? That is, taking the system to be an abstract process — a (possibly stochastic or nondeterministic) function of random variables — is the property balanced *in expectation*, or *in the limit*?

With this distinction in hand, we can ask: is the relevant property a property of the *system itself*, or of the *particular decisions* the system makes? Is satisfying the constraint required of the *process*, or of the *outcomes*?

Existing proposals for (and endorsements of) parity-based fairness constraints are not particularly careful about making these distinctions. On the first axis, most tend to phrase constraints as definitions, leading most naturally to an interpretation on which they make definitional (or constitutive) claims. Such claims, however, are implausible in the absence of stronger justification: the intuition that the constraints all have *something to do with* fairness is a far cry from an argument that they are *definitional* of fairness (if this wasn't already obvious from the aforementioned incompatibility results). As constraints that make good sense only for particular classes of systems, they lack the generality that definitions require. The statistical criterion embraced by a calibration constraint such as (11) above refers to the risk score. For systems that don't *use* risk scores, such as the general class of binary classifiers, it's not well-defined. So the criterion isn't generally applicable to binary classifiers. But fairness is a concept that *is* generally applicable to binary classifiers.[18] It follows that the calibration constraint does not define fairness. If the calibration constraint holds, it must be because it *follows from* the definition (or the nature) of fairness.[19] It seems more charitable, then, to attribute to these proposals the weaker claim that their preferred parity constraint articulates (non-definitional) necessary and sufficient conditions on fairness for particular subclasses of systems.

---

[17] More formally, we may mean by "statistical property" either a *(sample) statistic* — a function from sets of observations to numbers — or a *functional* — a function from statistical models to numbers. We may further associate with a functional an *estimator* — a (sample) statistic that (in some specified sense) approximates the value of the functional for "typical" observations. For the sake of readability, we will be loose with these distinctions; interested readers are referred to Wasserman (2004), chapters 6, 7, and 9.

[18] For the reader wondering whether there might be (relevant) binary classifiers to which fairness isn't applicable, a weaker premise will do: fairness applies to at least some systems that do not use risk scores.

[19] There is an analytic possibility that fairness is a disjunctive concept, defined along the lines of, "either calibration, if the system is a risk-score system, or equal accuracy, if the system is a binary classifier, or …." This is an unattractive option, because it seems that "fairness" refers to some more fundamental property: if we believe that calibration captures fairness for risk-scoring systems, we should believe this *because* calibration reflects the contours of a more fundamental notion of fairness when restricted to such systems. This point, however, is not essential here, since none of the extant proposals is even disjunctive.

With respect to the second axis (process *vs.* outcome), many authors articulate fairness "definitions" in terms of conditional independence properties; more generally, the proposals avail themselves freely of probabilistic notions. This is most naturally understood as putting constraints on the abstract, probabilistic system: fairness understood as a property of the process. Claims about the evaluability of the constraints in systems, however, are often phrased in terms of sample observations — which suggests an observations interpretation.[20]

Generally, most authors seem to presuppose *necessary and sufficient condition* views of fairness criteria linked with *system* interpretations. This linkage is sensible enough: thought of as necessary and sufficient conditions for a system to be fair, it is simplest to interpret them as conditions on the underlying system, rather than on the particular decisions that system makes in a particular case. The remaining interpretations can then be understood as derivative: assessing *observational* statistics on the particular decisions a system makes is a way of gathering *evidence* about the tendencies of the system's underlying process.[21]

So let us take this as a first interpretation for a parity constraint with respect to some property P: that a system probabilistically differs across racial groups with respect to P is necessary and sufficient for that system to be racially unfair; observations of a system's outputs which differ across racial groups with respect to P are derivatively relevant, because they provide evidence that the system is unfair in this way.

This is an especially natural stance for an individualist about race and racism. Recall that for the individualist, race is an inherent, non-relational property, and racism is a harm that is directed toward an individual. Viewed through this lens, a systematic violation of racial parity with respect to some metric P seems to be a violation of Aristotle's principle:the system subjects individuals with only morally *irrelevant* differences to morally *relevant* differences in treatment. To the individualist, it seems *procedurally unfair* to its decision subjects. In this light, it seems perfectly straightforward to think of a parity constraint violation as an instance of unfairness.

There are at least four independent reasons to be wary of this inference. First, it presupposes the moral irrelevance of race. And certainly, proponents of equity-based views such as those discussed below will resist this: race may be morally relevant because morally different treatment — such as compensation — is required.

Second, a violation of the Aristotelian Principle requires not just a lack of morally relevant differences in individuals, but also the presence of a morally relevant difference in treatment for such individuals. Even on the assumption that race is a morally irrelevant difference in decision subjects, for a parity constraint (on the interpretation here) to amount to a violation of the Aristotelian principle would entail that the

[20] This is not to say that authors are ignorant of the distinction. For example, Hardt *et al.* point out that subject to conditional independence constraints and assumptions of balance in the sample space, we can be assured of efficient estimation of the relevant joint distribution from samples.
[21] But see the end of this subsection for some alternative views.

purely probabilistic difference embedded in the algorithm is a morally relevant difference.[22] This might be resisted in a very general way: we might question whether probabilistic differences of this sort are morally relevant differences in *treatment* at all. They are not differences in the *decision outcome*, or the *consequences* (that is, the very things we most naturally think of as "treatment" differences): because these are stochastic systems, it is entirely possible that a system have a probabilistic difference in P with no corresponding difference in the outcomes, and *vice versa*. If the probabilistic difference is the necessary and sufficient condition for unfairness, then a use of a system violating P probabilistically is unfair, *even if* that use results in no difference in outcomes. The most intuitively plausible examples of this will be systems in which the outcome is *potentially* a moral harm or wrong; in such a case, the probabilistic difference in treatment amounts to subjecting an individual to a differential risk of harm. It is not, however, obvious that this is generally wrong. In US law, for instance, the general tort standard is one of "actual injury"[23] — one cannot usually sue for being subjected to a risk that does not materialize — and some have argued this follows from a logical incoherence in counting risks of wrongs as wrongs.[24]

Third, whether parity constraints are satisfied will, in general, be sensitive to how groups are chosen. In general, statistical property balance with respect to two independent properties does not guarantee balance with respect to combined properties. A system that is balanced across both Black and White subjects and across gender-conforming and gender-nonconforming subjects may nonetheless fail to be balanced between White gender-nonconforming and Black gender-conforming subjects. A property that is balanced across intersectional groups may likewise fail to have parity across the coarser groupings.[25]

Finally, there is the question we started with: just *which* (if any) of the statistical properties is morally relevant, and why? The various parity constraints that have been proposed suggest different — and, for many types of systems, inconsistent — properties. If, for risk score systems, a violation of a calibration condition constitutes a morally relevant difference in treatment, why does a violation of balance for false positives not? Scholars have only recently begun to take up this question in earnest; recent work by Rob

[22] It is tempting here to equate *probabilistic tendencies* of an algorithm with *dispositions*, and *dispositions* to *attitudinal bias*. This would be a mistake. The sense in which probabilistic tendencies are dispositions is a shallow one, akin to the sense in which electrons are disposed to repel one another. We take it that none of the proposals seriously countenances the view that machine systems have attitudes, or similar mental states.

[23] See, *e.g.*, Restatement (Second) of Torts §902

[24] See Heidi Hurd, 'The Deontology of Negligence', Boston University Law Review 76 (1996). Of course, this is a matter of great contention; the legal standard could well instead be justified for policy reasons, or simply wrong. We need not wade in here, but interested readers should consult Adriana Placani's (2017). "When the Risk of Harm Harms", Law and Philos 36; Stephen Perry's (2001) "Responsibility for Outcomes, Risk, and the Law of Torts", in Gerald Postema (ed.), Philosophy and the Law of Torts (Cambridge University Press, 2001); and Claire Finkelstein's (2003) "Is Risk a Harm?", 151 U. Pa. L. Rev. 963.

[25] See Kearns *et al.* (2018) for some examples. This phenomenon is a parallel to the Simpson-Yue paradox from elementary statistics, seen in the UC Berkeley gender bias case (Bickel et al. 1975).

Long[26] argues that false-positive-rate equality is sometimes inconsistent with fairness, and Brian Hedden[27] has argued that none of the proposed conditions except calibration is necessary for fairness. At the same time, David Grant[28] has argued that, in at least some cases, differences in calibration do not constitute morally significant differences in treatment.

Stepping back, it would be somewhat surprising if any of these criteria held generally, for a relatively simple reason: they are all insensitive to context, to the decision goals, and to the particular kinds of consequential harm and benefit that are at stake in a system's use. In the context of institutional justification, Tim Scanlon writes, "claims of procedural fairness … follow from particular institutional justifications, and the relevant standards of procedural fairness depend on the nature of this justification."[29] Just so for decision systems. What properties are, and are not, morally relevant, both for decision subjects and treatments, will depend upon what the system is doing, and why. It is a direct consequence of the impossibility theorems that *every* decision system will have differential treatment with respect to *some* properties that have *prima facie* moral relevance; the question for fairness is: for which of these properties, and when, is differential treatment justified in order to equalize treatment with respect to another?

There does not seem to be any reason to believe that there is a universal answer to this question. It seems clear that in some cases, accuracy matters morally: if a doctor is forced to choose between two different, equally risky treatments, being as accurate as possible seems morally required (all else being equal). In others, it seems that accuracy matters far less than avoiding false positives; the "beyond a reasonable doubt" standard in US criminal law reflects the judgment that a mistaken conviction is worse than a mistaken acquittal. So (to the extent that the wrongfulness of exposure to risk of the kind represented by a parity violation reflects the wrongfulness of the consequences of a statistical error) we should expect the moral importance of these risks to be contextual. If false-positive misclassification in a context constitutes a harm, then balancing such misclassification risk across groups matters morally (at least *prima facie*); in another context, it may not. Which statistical properties matter should — even on the probabilistic view — be a function of which consequences matter morally; since this will vary contextually, so too will the moral significance of the different kinds of risk embedded in the parity metrics.

Conflict between various parity metrics, then, seems to be a false paradox, so long as we recognize that statistical properties morally relevant in some contexts are not morally relevant in others. Likewise, general critiques of statistical fairness criteria show that such criteria cannot serve as analyses of (decontextualized) fairness, but do not establish that the criteria do not provide adequate tests in particular contexts. A given metric may or may not be morally significant in any given context; whether it is, is a question that must be resolved antecedently.

Before moving on from parity, it is worth acknowledging two variations on the claim made by the parity advocates. One separate individualist justification for parity constraints may come about via proxy

[26] Long 2020
[27] Hedden 2021
[28] Grant *ms*
[29] Scanlon 2018, p. 41

considerations: violations of (one) parity constraint are indications that there is a tacit proxy for race in the data. Such a justification is, we take it, not substantively different from those discussed in the previous section.

A second, more interesting, kind of answer centers group outcomes. On this view, it is not the (probabilistic) differential treatment of an individual that makes for unfairness, but the (probabilistic) expectation of differential outcomes for groups. This is a significant departure from the individualist account of racism discussed thus far: on this view, we need not think that any *particular individual* is harmed by a racially unfair decision system. However, it still derives its thrust from the Aristotelian principle: rather than requiring morally significant differences in treatment to be justified by morally significant differences at the individual level, we require it at the group level. This kind of desideratum is closely tied to egalitarian approaches to distributive justice. Here, the recipients of what is to be distributed are racial groups, and the particular parity criterion chosen corresponds to the "currency" of egalitarian justice, i.e., the unit that is to be distributed equally across the groups.[30] As such, it, like the individualist view considered in depth above, may still be expected to vary contextually (that is, with what is being distributed), and still relies on independent moral reasoning to identify which properties matter when distributed unequally.

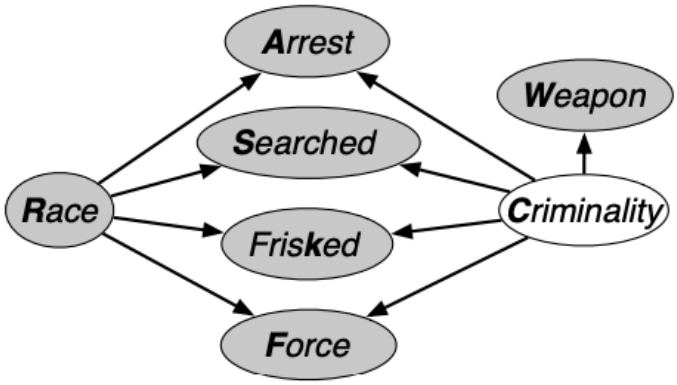
3C. *Race-Neutrality Approach III: Causal and Counterfactual Approaches*

Proponents of *causal* and *counterfactual* approaches to fair machine learning argue that core ethical concepts like racial discrimination are causal notions, and so reference to mere statistical or correlational criteria will never suffice to determine whether a system is discriminatory. Instead, the question of whether a given computational system is fair is intimately tied to questions around whether its output was subject to the right set of causal influences. Put roughly, a fair predictor must issue outcomes that are not inappropriately "caused" by race. As such, this set of methods draws from the broad literature on causal inference to devise its definitions of fairness.

Compared to the aforementioned purely statistical methods, casual approaches towards fairness are less able to be generally applied, because they work from a model of the causal relationships between race, other observed and unobserved attributes, and outcomes of interest. Heuristically, the causal model is meant to capture the *data-generating process*, depicting how the data that are inputs into the computational system are produced as a matter of how the world works. Causal models consist of a qualitative component—a diagram of nodes and edges, which represent the causal links among variables—and a quantitative component—a set of equations describing the functional relationships between a variable's value and those of the variables causally linked to it. This framework accounts for the causal effect of some variable of interest by *replacing* the "structural equation" for that variable with an equation that forces the target variable to be set to a particular value and then seeing how this change affects the values of variables causally downstream and most importantly, the value of the model's final outcome variable. Causal and counterfactual fairness proposals look to constrain the extent to which the value of the

[30] See Anderson (1999) for a detailed discussion.

system's output variable may change as a result of changes made to the race variable.[31] As an example, the following causal diagram was put forth by Kusner et al. 2018 to represent the network of causal relations that the authors believe to underlie the New York Police Department's Stop and Frisk dataset. Given this model, the authors estimate "counterfactual arrest rates," which are supposed to correspond to what arrest rates would have been "if every individual had been Black Hispanic [sic]."[32]



Of course, the theorizing required to posit a causal model of the data-generating process also introduces additional practical and theoretical challenges, since the same dataset and predictive task may be accompanied by different causal models.[33] Disagreement about the "correct" or "true" causal model presents a major practical problem for this set of approaches since, unsurprisingly, what a causal counterfactual approach calls for in order to achieve "algorithmic fairness" depends entirely on the model that is posited at the outset of the exercise. How to settle the problem of model ambiguity—figuring what the "right" causal model is— thus remains perhaps the most important task to be resolved for the framework. What is more, carrying out this task will be inherently value-laden, since different causal

[31] Influential works on causal and counterfactual approaches to algorithmic fairness include but are not limited to Kilbertus et al. 2017, Kusner et al. 2017, Russell et al. 2017, Nabi & Shpitser 2018, Chiappa 2019.

[32] Kusner et al. 2018, 18. It bears noting that, in regards to this causal diagram, the authors admit that, "We do not claim that this model has a solid theoretical basis, we use it below as an illustration on how to carry on an analysis of counterfactually fair decisions" (17).

[33] There are a number of reasons for disagreement about the "correct" causal model. Most notably, differences might stem from competing empirical theories of how variables are causally related to each other. But even those who agree on the empirics of the data generating process might put forth different causal diagrams because of differences modeling and abstraction choices. For example, the decisions regarding which causal relations are to be foregrounded in the diagram, and which are assumed in the background are choices about the level of granularity and abstraction of the model. Though this type of model ambiguity does not rest on empirical disagreement, it nevertheless makes a significant difference to what causal fairness requires. Lastly, it has also been argued that differences in causal theorizing about how social categories such as race and sex fit into the social wrold might also stem from differences in *normative* thinking about what these categories are and how they act causally in the world (Hu, forthcoming).

models emphasize different aspects of how the social world works and are thus better suited to illuminate different social explananda. For example, a causal model of the racial wage gap that treats as given, and thus leaves in the background of a diagram, racial segmentation of labor markets elides these structural features of employment as racialized and will be less able to point towards potential disruptions of these aspects of capitalism in a raced society. Decisions of what should be assumed in the background of a causal diagram and what should be explicitly represented as a causal factor in the foreground, what nodes and pathways are considered "racial" and which are not, are thus decisions that can only be made from a particular normative orientation towards matters of race and racial justice.

In contrast to the aforementioned purely statistical approaches, which are indifferent to one's background theory of race and its "causal effects," causal approaches to fairness require that theorists be forthcoming about the various social processes and mechanisms that they take to explain why the variables in a dataset take on the values that they do. That is, built into any application of the causal fairness framework is a theory of how the social world works, how the category of race works, and more broadly how social structural factors tie together different variables and produce the systematic correlations that machine learning systems tease out. In doing so, the causal modeling framework sets the stage for more sophisticated notions of fairness—ones that center racial categories as continually produced and reproduced by social processes and structures, which include the operation of predictive algorithms themselves.

In our view, the non-trivial task of theorizing about the causal structure of the social world and how race figures in that structure is indeed central to the task of building racially "fair" computational systems. Insofar as the concept of race and racial categorizations themselves materially affects people's lives, a good causal model recognizes the systematicity with which individuals different in "race" are different along many other of their "features" to arise out of the operation of social structures that create, maintain, and deepen racial differences across myriad social indices. This focus on *social* relations, processes, and structures as the key to understanding racial differences is shared by social-structural accounts of race.

And yet despite this shared orientation towards race as a social category that marks individuals for further social differentiation, prevailing causal and counterfactual approaches do not significantly depart from neither individualistic theories of race nor the Aristotelian principle. In fact, causal and counterfactual approaches see themselves as simply better equipped to achieve the Aristotelian principle. With a causal model in hand, they are more theoretically able to eliminate the "irrelevant" race factor from decision-making. Once again, the path towards treating likes alike reflects an individualist perspective on race and racism. The goal is to wash out the designated set of illicit race effects to make for "race neutralized" individuals or populations, so that the computational system cannot ensure that similar individuals receive similar treatment. Thus, the causal model is used to disclose what individuals are "really" like underneath the causal effects of their race. Beneath an individual's "race" are other of their qualities such as their underlying "merit," which are considered the proper basis of treatment. Conceived of in this way, race is a feature that hinders the system's ability to access these non-raced, "truer" qualities of groups. Race is a social category that the computational system should see through or past to avoid unfairness.

Prevailing causal and counterfactual approaches, like the data- and parity-based methods, therefore see racial justice as achievable if only computational systems can be designed to avoid drawing on racial

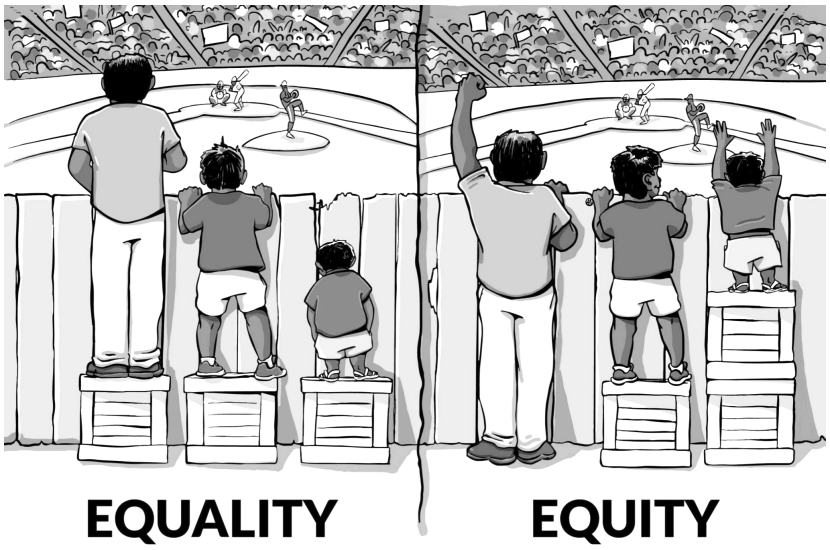
information or otherwise eliminate the distortionary effect of racial membership on the system's computations thinking. Nevertheless, we agree that a framework that elucidates how the concept of race generates material differences in a society presents the right starting point from which we ought to theorize racial fairness. But importantly, careful theorizing about the role of race in the causal structure of the social world should not only guide towards washing out effects of race on other features in a dataset. As social-structural accounts of race emphasize, a picture of race as causally efficacious stands central in a theory of race that sees racialization as a social process that is both historical and ongoing. Insofar as the systems we build are to be deployed in a society characterized by racial stratification, there is no sense in which computational systems can function "neutrally" with respect to race. The background facts of racial injustice which reign make it the case that computational systems embedded in key social institutions are already implicated in processes that either serve to sustain the system of racial hierarchy or actively work to undermine it. This perspective then suggests that computational systems be built not just to achieve some standard of procedural fairness in the form of not basing any of its operation on racial information or otherwise eliminating any effect of race. Instead, this analysis suggests that approaches to racial justice in these systems must make reference to some end state of racial inequality. Methods toward fairness earn their keep by reference to the patterning of racial egalitarianism achieved and the extent to which the system of race is weakened in the society.

4. How *Ought* We Address Racism: Normative Considerations

In the previous section, we argued that despite the diversity of frameworks that have been proposed in the literature on achieving racial fairness in computational systems—ones based on transforming the data inputs into the systems, placing statistical constraints on the behavior of the systems, and building systems that limit "causal effects" of race on their outputs—all of them, in one way or another, proceed from an individualistic conception of race and an Aristotelian conception of fairness. In this section, we explore ways of shifting our presuppositions about racism and fairness, with hopes that these conceptual changes may shed light on how to move forward in the debate on fairness in computational systems.

We saw above that the normative principle guiding much of the scholarly work in this area is the Aristotelian idea that we should treat "likes alike, and unlikes unlike." According to this principle, equality's main concern is that morally-irrelevant factors are not the basis for differential treatment. There are two ways to apply this conception of equality when injustice is at issue: one can either *ignore* the morally-irrelevant factors, or *compensate* for their effects. In the context of gender equality, the former approach tends to align with the doctrine of gender neutrality—an approach that emphasizes *sameness* across genders—according to which women are the same as men and thus should be treated in roughly identical ways. The latter approach tends to align with the doctrine of special benefits—an approach that emphasizes gender *difference*—according to which women and men can only be treated alike by engaging some compensatory mechanism to make up for the biological differences, or the historical and existing disadvantages that women have compared to men.[34] These two approaches are often referred to as an "equality" approach (sameness of treatment), and an "equity" approach (address differences to get sameness of outcome) approach, illustrated in a widely circulated image that captures the difference:

[34] For a more extensive discussion of the Aristotelian principle and a critique of the difference/sameness approaches, see Catharine MacKinnon (1987).



Notice, for example, that if the differences between the three audience members had no effect on whether they could see the game—imagine an image with three kids who each needed one box see over the fence, and also had differences in hair color—then the move from equality to equity would not be necessary *because the hair color difference among the kids is irrelevant to what matters in the situation; mere* differences don't matter. All that would matter in that case is that the three kids each gets a box to stand on to allow them to see over the fence. The equity model—so differential treatment—is required in the case pictured to deal with what matters morally there: whether one can see, and so enjoy, the game.

We can see the Aristotelian principle at work throughout the contemporary debate about racial fairness in machine learning. That the field has largely embraced these two conceptions of equality is evident in the approaches to fairness surveyed in this essay. Quite literally, the fairness through unawareness approach tries to eliminate the effect of (morally-irrelevant) race on the system's decisions by ignoring it altogether. The data-driven strategies in general embody the impulse to ignore race by eliminating its influence on the decisions of the system. Likewise for the causal and counterfactual approaches, which attempt to compensate for (or perhaps eliminate) the effects of race so that the computational system can treat individuals across races alike. The statistical parity methods are an effort to achieve equality through imposition of parity constraints, which are intended to achieve some kind of balance in the way individuals are treated across races.

It bears noting at this point that although these proposals are united in presupposing the Aristotelian principle and individualistic conceptions of race and racism, they are still able to reach different conclusions about how to address racial injustice in algorithms. This is because they represent different views about what count as morally relevant similarities, both in individuals and treatments. Causal approaches differ from data-driven neutrality approaches in taking causal sequelae of race, in addition to race itself, to be morally irrelevant. Similarly, different parity approaches take different statistical properties to be morally relevant. One might thus see this built-in versatility of the Aristotelian principle to present a good reason for its adoption as a guide toward racial fairness. After all, how could a principle

that calls for systems to "treat likes alike and unlikes unlike" lead us astray? We will raise two concerns here.

First, the fact that a wide diversity of approaches seem to adhere well to the principle highlights one significant shortcoming of the Aristotelian principle as a basis for designing approaches to algorithmic justice. There is a substantive set of questions that must be answered before applying the Aristotelian principle: What features are or are not morally relevant? In which respects must individuals be similar, and in which respects must treatments be similar? Until these questions are answered, the principle is underdetermined. But, of course, answers to these questions require *moral* reasoning, not *technical* reasoning. Small wonder, then, that efforts to reach technical solutions to algorithmic fairness have fallen flat.

The key question of what should be taken to be morally relevant similarities across individuals is central in both data-based neutrality as well as causal and counterfactual approaches. For both methods require decisions to be made about whether a feature's relationship to race, either causally or merely correlationally, warrants its exclusion from the dataset or its effects to be washed out entirely or whether the feature is nevertheless a sound basis for differential treatment. And as discussed in the preceding section on parity-based approaches to fairness, in light of the "impossibility of fairness" mathematical result, the central debate among the different opposing camps is precisely the question of what statistical metric should be taken to correspond to what is morally relevant in treatment. Does similar treatment across races mean achieving similar *accuracy* rates across racial groups or similar *positive classification* rates or similar *false positive* rates?

There is no way to enter into, let alone resolve, these debates without weighing in substantively on normative matters. Scientific inquiry often tries to avoid drawing on moral and political values in order to be "objective." We do not need to get into larger discussions of scientific objectivity here, however.[35] We are assuming that machine learning is part of a decision-making process: it selects what information is or is not relevant for the decision, policy, or action at hand. Decision making and the resulting actions, especially when they concern the distribution of key social benefits and burdens, are rightly subject to moral considerations.

But how do we proceed if we need to bring values and morality into the discussion? Aren't values "subjective" or "culturally relative" and so not something that can play a substantive and legitimate role in machine learning? The first thing to note is that our discussion thus far demonstrates that values are already playing a role and furthermore, *should* play a role. There is no way to opt out of weighing in substantively in a way that implicates one's values, since to do nothing and continue with the status quo ante is just as value-laden as intervening in reigning circumstances and making a change. What is more, the suggestion that values are "subjective" as in "just a matter of taste" (you like chocolate; I like vanilla) is not plausible. We can allow that values are contextually variable, and that there is disagreement about

[35] There is a long tradition in feminist epistemology and philosophy of science calling the ideal of objectivity into question; see, *e.g.*, Longino (1990) and Douglas (2000).

values, but still insist that there be disciplined and systematic inquiry into the values that currently play a role, and what values would be better at this point in time in this context.[36]

Philosophers (and others) have been concerned to articulate and defend moral principles in a systematic way for thousands of years. We will not defend a particular set of principles or overarching theory that identifies what is morally relevant, but here we present a few examples of approaches to give a sense of what might be at issue:[37]

> **Luck Egalitarianism:** "[J]ustice requires that no-one should be disadvantaged relative to others on account of 'brute' bad luck, whereas inequalities that arise through the exercise of personal responsibility are permissible." (Miller 2017).

A luck-egalitarian would consider it morally relevant whether a feature is due to bad luck or choice. Are X and Y different because X was lucky and/or Y was not? If so, we should make an effort to compensate Y for bad luck, e.g., an accident that rendered them disabled, or, even, being born a woman, or Black. In contrast, if the difference is due to a deliberate choice in which luck played no role, there may be no need to compensate for the difference. It remains, however, a significant hurdle in luck-egalitarian accounts of justice to come to an account of "luck" and "personal free choice" that makes for a plausible theory (Lippert-Rasmussen 2018).

> **Preference Utilitarianism:** The good consists in the satisfaction of preferences and the bad the frustration of preferences; so one ought morally to maximize the satisfaction of preferences.[38]

A preference utilitarian would consider the only morally relevant consideration in making a decision to be whether it maximizes the satisfaction of (subjective) preferences. So the only morally relevant features of individuals are what they prefer (not their race, gender, etc.). All preferences matter equally, no matter how ignorant, wicked, trivial, and so the view can be highly counter-intuitive.[39]

---

[36] There is, as you might expect, a huge literature in moral philosophy on issues of moral disagreement, moral relativism, and moral subjectivism. We are not in a position to address the concerns fully in this paper. However, for a start, see Midgley (1981/2003) and Shafer-Landau (2012).

[37] Philosophers distinguish "historical principles" from "end-state principles" of justice. Historical principles focus on the procedures that led to the present distribution: how decisions were made, what factors were taken into account or ignored, what was treated as morally relevant, were transfers carried out in a way that respected the principles of proper exchange? End-state principles, in contrast, focus on the pattern of distribution that results as such: does applying the principle result in a distributive pattern that maximizes goodness? Or does applying the principle result in a pattern that deviates from equality only in ways that are morally acceptable? Of the principles mentioned here, the first is a historical or procedural principle and the latter two are end-state. Some theories of justice combine both kinds of principles.

[38] Preference utilitarianism is one of a broad range of *consequentialist* moral theories. For a description and discussion of consequentialist accounts, see, e.g., Sinnott-Armstrong 2019.

[39] Another serious problem with preference satisfaction theories is the phenomenon of adaptive preferences. Under conditions of oppression, individuals' preferences might adjust to what is deemed

> **Democratic Egalitarianism:** A society is just if and only if its members stand in relations of equality, and there is no oppression, i.e., there are no laws, policies, norms, and such that deprive individuals of equal moral standing. To achieve justice, oppressive structures should be dismantled. (Anderson 1999; Scheffler 2010; Arneson 2013)

Democratic egalitarians are concerned with what is called *relational equality* and the structures that mediate our interactions. The project of justice is not to produce a society in which we don't suffer the effects of bad luck, or to produce a society in which as many of us as possible get whatever it is we want. The project is to eliminate relations of domination based on unearned status. Notice that this is a structural approach to justice and differs from the interactionist account of individual racism: what counts as morally relevant for consideration by the Aristotelian principle includes social position and not just intrinsic features of individuals. The view faces some challenges in determining what is "unearned" and how to circumscribe the scope of the view.

We are not arguing that the Aristotelian principle is wrong, or misguided. Rather, we maintain that it, alone, cannot do the necessary work of determining what treatment is fair and just. The principle may be appealingly "obvious" and reliance on it may seem to be "value neutral". But in fact, it is effective as a normative tool only if it is implicitly or explicitly augmented by more substantive, and potentially controversial, accounts of moral relevance. Those concerned with racial justice should not shy away from substantive moral claims, but such claims should be supported by disciplined normative inquiry.

Let's turn now to a second shortcoming of the Aristotelian principle. Our discussion thus far has demonstrated that many morally relevant features are highly contextual, diachronic, mutable, or relational; this is to be expected in the particular context of algorithmic racism, because the mechanisms of racism and white supremacy are, likewise, highly contextual, diachronic, mutable, and relational. So we should expect there to be many contexts in which the features relevant to applying the Aristotelian principle are subtle. In machine learning contexts, however, such subtle features are virtually never represented directly as features in the data; the differences and similarities to which most algorithms (and their designers) have direct access in the form of concretely-defined collected features in the dataset either are, or are assumed to be, invariant, stable, and intrinsic. So, in contexts in which the morally relevant features are subtle, using features accessible in data for making *moral* similarity judgments is a straightforward mistake. This mistake becomes even more serious when the moral similarity judgment becomes codified in the form of a statistical metric or technical constraint that is applied broadly to all computational systems of some kind. This kind of tendency to universalize certain ways of operationalizing ethical principles goes directly against the context-sensitivity of good moral theorizing.

---

possible or realistic, or they might even take on preferences that run counter to their flourishing. A preference utilitarian who aims to satisfy actual preferences might leave people with only what they would hope to get under conditions of injustice or worse, might facilitate the satisfaction of preferences that are in fact harmful to objective markers of well-being. This is not justice. See Nussbaum (2001); Khader (2009).

So not only does use of the Aristotelian principle require substantive normative judgments about moral relevance, but also, because such judgments are likely to be highly context-specific, there can be no one-size-fits-all technical solution to addressing racism in algorithms. Any "solution" to the fairness problem must be sensitive to the particular context in which an algorithm is used: the sociotechnical systems in which the algorithm is embedded, the data on which it relies, the specific mechanisms by which power is distributed (and White supremacy is enforced) in systems the algorithm interacts with, and the particular kinds of values and harms at stake. Because the details of systemic racism vary wildly across contexts—compare housing, to medicine, to public service provision—we should expect the details of building algorithms that resist systemic racism to vary as well.

For example, the interpretation and interpolation of raw data, the statement of optimization goals and constraints in numerical terms, and an assessment of what constitutes a system's success uncontroversially involve contextual considerations. Whether we should discard data with missing features depends upon whether we think that data is importantly informative; whether we should weigh false positives and false negatives equally in accuracy metrics depends upon whether the two kinds of mistake are equally bad given our purposes. And even in deciding what the task of treating likes alike itself requires, there is an ineluctable contextual element. It should not surprise us, then, that whether parity of one sort or another is an appropriate metric in a given context will depend upon what the outcomes are, and, more specifically, what moral values are implicated in them.

Where do we go from here? We have identified two themes running throughout the predominant frameworks toward racial fairness in machine learning thus far: individualistic conceptions of race and racism and Aristotelian conceptions of equality. To chart a path forward in the debate, we propose that we part with these conceptualizations of race, racism, and equality. Luckily, there are alternatives available to us.

A more plausible picture of race and racism acknowledges the fact that race is a product of and inextricable from the social world in which it exists, that it is enacted and reproduced by systems of social relations, institutional structures, and material conditions, while simultaneously interacting with and shaping these same features of the social world. Such a structural view naturally lends itself to an account of racism that foregrounds analyses of power and domination in interacting systems, and approaches to anti-racism that focus on dismantling the institutional scaffolds that produce, enforce, and recreate white supremacy.

In contrast to Individualist Race, we suggest a social-structural conception of race (see section 2, "Race as Relational"):

> **Social Structural Race:** races are positions in a social structure that distributes power to individuals or groups in virtue of their occupying a node in the structure. The differences that we see between races are consequences of efforts to create, maintain, and reinforce the system and the unequal distribution of power.

Moreover, injustice is not only a matter of wrongful or harmful treatment of individuals by individuals but also lies in the relations that make up a social structure.

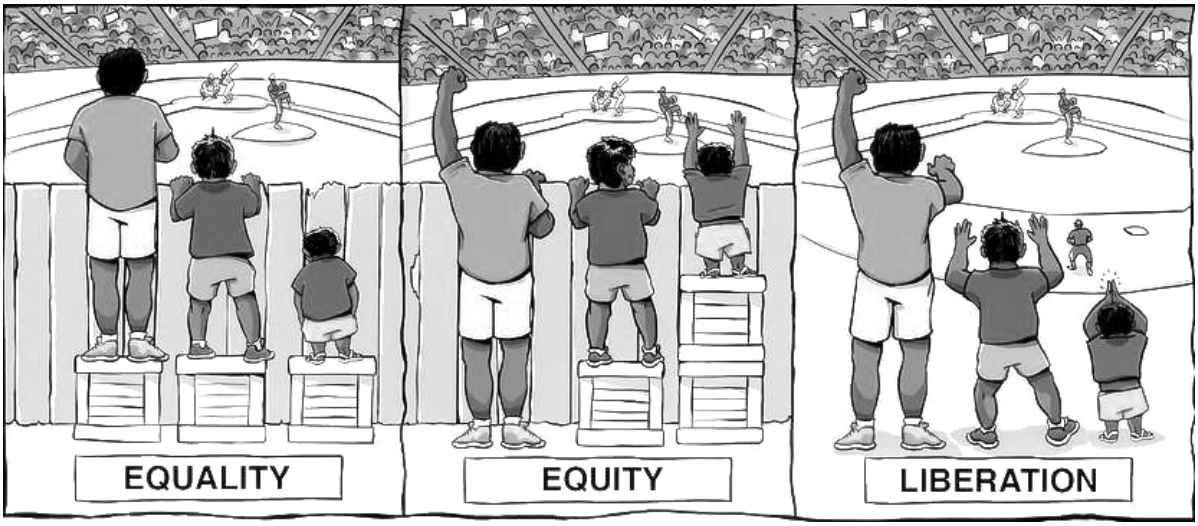
> **Structural Racism**: a system or a structure is racist when the network of relations that compose it distributes goods—not only material goods but also immaterial goods such as power, status, security, and opportunity—or misinterprets rights, in ways that unjustly advantage or disadvantage one or more races.

For example, attending to structural racism in the criminal justice system would locate the problem not only in individual police officers or judges, the "bad apples," but in, say, the power and authority distributed to police to "legitimately" use lethal force when the officer has "reasonable suspicion" that the subject poses a significant threat of serious bodily injury or death to themselves or others (CFR § 1047.7). A system that distributes power in this way may be wrong because of background conditions, including the culture and training of police officers, how the standard of "reasonable suspicion" is understood, the history of race relations and consequent racial stereotyping, and the availability of other mechanisms to prevent danger, injury or death. Even if no police officer has so far used lethal force and so there is no instance of wrongful racial "treatment," these conditions may suffice to count as perpetrating structural racism.

These structural alternatives present us with a challenge. On the one hand, it would seem that because racial differences are the result of background injustice, a compensatory approach is warranted. This is a strategy found in traditional justifications for affirmative action and in the shift from equality to equity (recall the image above). For example, African-Americans suffer from discrimination (both individual and institutional), so we should make special efforts to include and protect them in meaningful ways. And yet, on the other hand, if the system itself is unjust, should we really be seeking ways to promote the inclusion of those who are disadvantaged in the system? That is, is the right approach to simply try to give them a fairer chance or a leg up in the same system of structural inequality? Shouldn't we instead be asking: How and why does the system create positional vulnerability and disadvantage? Rather than simply leaving the system as it is and facilitating inclusion within it, shouldn't we also work to change or disrupt the system itself?

Let's return again to the (relatively simple) example of a university admissions process. Whereas a crude "treat likes alike" approach maintains that fairness requires a computational system to treat all students who score a 1200 on the SAT the same, regardless of their race, a difference-based approach might recognize that because of differences in the distribution of SAT scores across racial groups tracks morally irrelevant differences (perhaps bad luck), a Black student with a 1200 score should be treated like a White student with a 1300 score. In contrast, a structural approach would see that different observed distributions in SAT scores are not just examples of "racial difference" that can be resolved simply by instituting algorithmic affirmative action. They are the result of historical as well as ongoing anti-Black racism that has been built into the educational system. To affirm them as mere "differences" is to leave out of the story social processes of racialization and racial domination as continually reproducing and sustaining this "difference"—ongoing social processes of racialization that distribute power among individuals and groups of which computing systems increasingly play a part. The pertinent question of justice then becomes: why are we employing standardized tests in the first place? Who *should* be admitted to universities? What is the proper function of higher education in a society? *How has race and racism already affected our ordinary answers to these questions?* Should we do things differently? These are questions that an approach such as Democratic Egalitarianism and other structural approaches to justice can help us answer.

This alternative conceptualization of the problem of social injustice is well captured in an additional panel to the familiar image of equality and equity presented earlier—one that points to the *social basis* for the disadvantages that result from height differences in being able to enjoy the baseball game: the fence. The image does not capture all of the ways that the equality/equity model is inadequate, but it is a start.



If similarities and differences between races (like height) are just "the facts" that have to be taken into account as we decide how to achieve fairness, and the social system is taken as given (being positioned behind the fence), then the solution is framed around identifying and addressing the important facts that will make the system work effectively. (How many boxes do we provide so everyone can see?). Who is entitled to inclusion and who not? How do we include them? It would also seem that all we need in order to address the problem is a broad commitment to fairness; no other values are at issue, because the status quo system is just what it is, a *neutral* feature of our social world.

But if the system is unjust and misrepresents disadvantages as natural, intrinsic, "given," then the status quo is already causing disadvantage—it provides benefits to some and harm or risk of harm to others. Taking the system to be "neutral" suggests that any deviation from it requires special justification. But to continue with the system as-is requires justification. As a result, those who focus on sameness and difference are in danger of being complicit in injustice: the sameness/difference model naturalizes the unjust system that *produces* difference. But once we recognize that we are the system that creates race and racial hierarchy and that we create and maintain the fence, we can properly turn our attention to dismantling the underlying social processes and structures that constitute the unjust system.[40]

[40] It bears noting that the three-paneled image continues to leave out a dimension of the problem that we consider important: not only are the disadvantages and advantages created by the system—whether or not one can see the game is a result of the fence being there, a non-natural structure erected by social processes—but in many cases, the relevant *differences* among individuals and groups are themselves generated by the system. That is, in visually representing advantage and disadvantage via height, the image naturalizes differences among individuals. It does not suggest that the fence is what causes the differences in height. The SAT example demonstrates the point: as a matter of fact, the differences in

5. Conclusion

Let's return to the principles introducing the issues of race and racism in Section 2. We have argued that many of the efforts within AI to take race and racism into account presuppose that race is an intrinsic feature of individuals (*Individualist Race*). Moreover, the interventions to address potential racism relies on the background normative claim that we should "treat likes alike and unlikes unlike" (the *Aristotelian Principle*). By bringing these two ideas together, current approaches in AI assume that computational systems are anti-racist if they treat individuals of different races alike when they are like and unlike when they are unlike (Countering *Individualist Racism*).

We have argued this picture does not capture the problem of systemic racism. Even if there are forms of racism that are a matter of discrimination against individuals, these are not the only forms that anti-racists must address. A better account of race maintains that race is an imposed social status that is a matter of one's position in a social structure (*Structural Race*), and racial injustice lies not only in our interactions but in the background relations and social constraints (laws, norms, material conditions) that structure these interactions (*Structural Racism)*. According to this structural approach, the Aristotelian principle—without further normative principles—is insufficient to respond to the process of racialization, because it is silent about what differences are morally relevant (*Moral Relevance*). The principle also obscures the ways in which many of the observed "differences" among individuals are themselves produced by systems of oppression. This means that a system's "treatment" of individuals is also situated within institutions and structures that we cannot assume are just. We must not only be attentive to the information provided to facilitate decisions, but why these decisions, these policies, these institutions, have the power they do. Without this broader critical perspective, existing efforts to avoid systemic racism in data-based computational systems will not address systemic racism, and will likely maintain it.

This is not to say that the approaches discussed in Section 3 can never be useful tools to addressing racism. They can be, even while adopting a structural account of race and racism. However, we must recast them: they are not analyses of *what it is* for an algorithm to be racially biased, or white supremacist; they are, rather, techniques that can serve as useful tools, in particular contexts, for addressing racism embedded in algorithmic systems. Using the tools effectively requires consultation with domain experts who can do the work of racial justice analysis in their relevant area of expertise, as well as with advocates, organizers, and critics. There is a tremendous amount of humanist and social-scientific work that has helped us to understand the historical context and processes of racial formation; there is also a tremendous amount of work on justice, and more specifically racial justice, in virtually every arena of social life. Our recommendation is that MIT should form a genuinely interdisciplinary institute in SHASS

---

vocabulary between Black and Brown students and White students is not a "mere difference," a feature of their background circumstances that the SAT mistakenly tracks. The differences are the result of historical and ongoing racism in the education system that continues to affect life chances, even when affirmative action measures are implemented and even if we "take down the fence." Extensive damage has been done, and the problem won't be solved just by changing one institution or one law (etc.). A more radical systemic overhaul is needed. This is certainly not the job of solely engineers or computer scientists, but they must be part of the broader movement that does it.

that promotes research across disciplines but also includes engagement with activists, stakeholders (including those directly impacted), and those representing state and corporate interests.

Works Cited

Adebayo, Julius and Lalana Kagal. (2016). "Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models." arXiv preprint arXiv:1611.04967.

Alcoff, L. M. (2000). "Is Latina/o Identity a Racial Identity?" In Jorge J. E. Gracia and Pablo De Grieff (eds) *Hispanics/Latinos in the United States: Ethnicity, Race, and Rights*, pp. 23-44.

Anderson, E., 1999, "What is the Point of Equality?", *Ethics*, 109: 287–337.

——, 2010. "The Fundamental Disagreement between Luck Egalitarians and Relational Egalitarians", *Canadian Journal of Philosophy* (Supplementary Volume), 36: 1–23.

——, 2012. "Equality", in D. Estlund (ed.), *The Oxford Handbook of Political Philosophy*, Oxford: Oxford University Press, pp. 40–57.

Appiah, Kwame Anthony. 1996. "Race, Culture, Identity: Misunderstood Connections." In *Color Conscious*, edited by Amy Gutman and K. Anthony Appiah. Princeton: Princeton University Press, pp. 30ff.

Arneson, Richard. 2013. "Egalitarianism", *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2013/entries/egalitarianism/.

Barocas, Solon & Andrew Selbst. (2016). "Big Data's Disparate Impact." *California Law Review* 104: 671.

Bolnick, D. A. (2008). Individual ancestry inference and the reification of race as a biological phenomenon. In B. A. Koenig, S. S.-J. Lee, & S. S. Richardson (Eds.), Revisiting race in a genomic age (pp. 70-85). New Brunswick, NJ: Rutgers University Press.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016, December). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In *Proceedings of the 30th International Conference on Neural Information Processing Systems*: 4356-4364.

Blum, Lawrence. (2002). *"I'm Not a Racist, But . . ." The Moral Quandary of Race.* Ithaca, NY: Cornell University Press.

Chiappa, Silvia. (2019). "Path-Specific Counterfactual Fairness." *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1): 7801-7808.

Chouldechova, Alexandra & Aaron Roth. (2020). "A Snapshot of the Frontiers of Fairness in Machine Learning." *Communications of the ACM* 63(5): 82-89.

Corbett-Davies, Sam & Sharad Goel. (2018). "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv preprint arXiv:1808.00023*.

Douglas, Heather. "Inductive Risk and Values in Science." In *Philosophy of Science* 67.

Espiritu, Yen Le. (1992). "Ethnicity and Panethnicity." In *Asian American Panethnicity,* Temple University Press, pp. 1-18.

Gajane, P. & Pechenizkiy, M. (2017). On Formalizing Fairness in Prediction with Machine Learning. *arXiv preprint arXiv:1710.03184*.

Glasgow, Joshua. (2009). *A Theory of Race.* New York Routledge.

Glasgow, Joshua, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer. (2019). *What is Race? Four Philosophical Views.* New York: Oxford University Press.

Gooding-Williams, Robert. (1998). "Race, Multiculturalism, and Democracy." *Constellations* 5:18-41.

Grant, David. (2021). "Fairness in Prediction." *manuscript* [presented at Jain Family Institute workshop in digital ethics].

Hardimon, Michael. (2017). *Rethinking Race: The Case for Deflationary Realism.* Cambridge, MA: Harvard University Press.

_____. (2019). "Four Ways of Thinking About Race." *The Harvard Review of Philosophy* 26: 103-113.

Hardt, Moritz, Price, Eric, and Srebro, Nathan. (2016). "Equality of Opportunity in Supervised Learning." *Advances in Neural Information Processing Systems.* https://arxiv.org/abs/1610.02413

Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Noûs* 34(1): 31-55.

Hedden, Brian (2021). "On Statistical Criteria of Algorithmic Fairness." In *Philosophy and Public Affairs* 49: 209-231. https://doi.org/10.1111/papa.12189.

Hellman, Deborah. (2020). "Measuring Algorithmic Fairness." In *Virginia Law Review* 106: 811.

Hu, Lily. (2021). "Causation in the Social World." *Manuscript.*

Huq, Aziz H. (2019). "Racial Equity in Algorithmic Criminal Justice." In *Duke Law Journal* 68: 1043.

Khader, S. J., 2009, "Adaptive Preferences and Procedural Autonomy," *Journal of Human Development and Capabilities*, 10: 169–187.

Kilbertus, Niki, Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). "Avoiding Discrimination Through Causal Reasoning." *Proceedings of the 31st International Conference on Neural Information Processing Systems*: 656-666.

Kim, Pauline T. (2017). "Data-Driven Discrimination At Work." *William & Mary Law Review* 58: 857.

Kim, Won, Byung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, & Doheon Lee. (2003) "A Taxonomy of Dirty Data." *Data Mining and Knowledge Discovery* 7(1):81-99

Kroll, Joshua A., Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. "Accountable Algorithms." *University of Pennsylvania Law Review* 165: 633.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Proceedings of the 31st International Conference on Neural Information Processing Systems*: 4069-4079.

Lippert-Rasmussen, Kasper, "Justice and Bad Luck", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2018/entries/justice-bad-luck.

Long, Robert. (2020). "Fairness in machine learning: against false positive rate equality as a measure of fairness". *arXiv preprint* https://arxiv.org/abs/2007.02890.

Longino, Helen. (1990). *Science as Social Knowledge*. Princeton, Princeton University Press.

Mallon, Ron. (2006). "'Race': Normative, Not Metaphysical or Semantic." *Ethics* 116(3): 525-551.

Mayson, Sandra G. (2019). "Bias In, Bias Out." *Yale Law Journal 128*: 2218.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). "A Survey on Bias and Fairness in Machine Learning." *arXiv preprint arXiv:1908.09635*.

Midgley, Mary. 1981/2003. "Trying Out One's New Sword." *Heart and Mind*. New York: Routledge, pp. 80-87. http://www.christopherlay.com/MidgleySword.pdf

Miller, David, "Justice", *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/justice/>.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, & Kristian Lum. (2018). "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions." *arXiv preprint arXiv:1811:07867*

Narayanan, Arvind. (2018). Translation Tutorial: 21 Fairness Definitions and Their Politics. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency.*

Nussbaum, M., 2001, "Adaptive Preferences and Women's Options," *Economics and Philosophy*, 17: 67–88.

Omi, Michael & Howard Winant. (1994). *Racial Formation in the United States from the 1960s to the 1990s.* New York: Routledge.

Outlaw, Lucius. (1996). *On Race and Philosophy.* New York: Routledge.

Richardson, Rashida, Jason Schultz & Kate Crawford. (2019). "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice." *NYU Law Review* 94: 15

Nabi, Razieh & Shpitser, Ilya. (2018). "Fair Inference on Outcomes." *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*: 1931-1940.

Risch, N., Burchard, E., Ziv, E., & Tang, H. (2002). "Categorization of humans in biomedical research: Genes, race, and disease." Genome Biology, 3, 2007.1-2007.12.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L.A., et al. (2002). "Genetic structure of human populations." Science, 298, 2381-2385.

Russell, Chris, Kusner, M. J., Loftus, J. R., & Silva, R. (2017). "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness." *Proceedings of the 31st International Conference on Neural Information Processing Systems*: 6417-6426.

Sattigeri, Prasanna, S. C. Hoffman, V. Chenthamarakshan, & K. R. Varshney. (2018) "Fairness

GAN" arXiv preprint arXiv:1805.09910.

Scanlon, T. M.. 2018. *Why Does Inequality Matter?* (Uehiro Series in Practical Ethics). OUP Oxford. Kindle Edition.

Scheffler, Samuel. 2010. *Equality and Tradition: Questions of Value in Moral and Political Theory*, Oxford and New York, Oxford University Press.

Shafer-Landau, Russ. 2012. "Ethical Subjectivism." In Russ Shafer-Landau and Joel Feinberg (ed.), *Reason and Responsibility,* 11th edition. Boston: Cengage Learning, pp. 535-546.

Simoiu, Carmelia, Sam Corbett-Davies, and Sharad Goel. (2017). "The Problem of Infra-marginality in Outcome Tests for Discrimination." https://arxiv.org/pdf/1607.05376.pdf.

Spencer, Quayshawn. (2014). "A Radical Solution to the Race Problem." *Philosophy of Science* 81(5): 1025-1038.

_____. (2015). "Philosophy of Race Meets Population Genetics." *Studies in*

History and Philosophy of Biological and Biomedical Sciences 52: 46-55.

Stocking, George. (1994). "The Turn-of-the-Century Concept of Race." *Modernism/ Modernity* 1(1): 4-16.

Wasserman, Larry. (2004). *All of Statistics: A concise course in statistical inference*. New York, Springer.

Yang, Crystal S. & Will Dobbie. (2020). Equal protection under algorithms: A new statistical and legal framework. *Mich. L. Rev.*, *119*, 291.

Zemel, Richard, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. (2013). "Learning Fair Representations." *International Conference on Machine Learning*: 325-333.

Zheng, Robin. (2018). Review of Michael Hardimon *Rethinking Race: The Case for Deflationary Realism.* https://ndpr.nd.edu/news/rethinking-race-the-case-for-deflationary-realism/